Hindawi
Security and Communication Networks
Volume 2021, Article ID 5552501, 11 pages
https://doi.org/10.1155/2021/5552501

Research Article

PurExt: Automated Extraction of the Purpose-Aware Rule from the Natural Language Privacy Policy in IoT

Lu Yang,[1,2] **Xingshu Chen**,[2,3] **Yonggang Luo**,[2,3] **Xiao Lan**,[2,3] **and Li Chen**[1,2]

[1]*College of Computer Science, Sichuan University, Chengdu 610065, China*
[2]*Cyber Science Research Institute, Sichuan University, Chengdu 610065, China*
[3]*College of Cyber Science and Engineering, Sichuan University, Chengdu 610065, China*

Correspondence should be addressed to Yonggang Luo; iamlyg98@scu.edu.cn

Received 15 February 2021; Revised 14 April 2021; Accepted 24 April 2021; Published 8 May 2021

Academic Editor: Ahmed Meddahi

The extensive data collection performed by the Internet of Things (IoT) devices can put users at risk of data leakage. Consequently, IoT vendors are legally obliged to provide privacy policies to declare the scope and purpose of the data collection. However, complex and lengthy privacy policies are unfriendly to users, and the lack of a machine-readable format makes it difficult to check policy compliance automatically. To solve these problems, we first put forward a purpose-aware rule to formalize the purpose-driven data collection or use statement. Then, a novel approach to identify the rule from natural language privacy policies is proposed. To address the issue of diversity of purpose expression, we present the concepts of explicit and implicit purpose, which enable using the syntactic and semantic analyses to extract purposes in different sentences. Finally, the domain adaption method is applied to the semantic role labeling (SRL) model to improve the efficiency of purpose extraction. The experiments that are conducted on the manually annotated dataset demonstrate that this approach can extract purpose-aware rules from the privacy policies with a high recall rate of 91%. The implicit purpose extraction of the adapted model significantly improves the F1-score by 11%.

1. Introduction

IoT applications cover all aspects of people's lives, such as smart homes, wearable devices for health management, and traffic monitoring. While enjoying the convenience brought by IoT applications, users also bear a greater risk of personal information leakage than general applications. The data collected by IoT applications comes not only from smartphones but also from various smart devices that are closer to users' daily lives, so the data is more sensitive [1]. The collection and use of data should be more strictly in compliance with regulations. Furthermore, because of the potentially unobtrusive nature of IoT data collection, users may not be aware of what information is collected and why it is collected [2, 3]. To solve the above issues, data protection laws and regulations, such as the General Data Protection Regulation (GDPR), require that before collecting and using data, IoT application providers must inform users of the privacy policies and obtain their consent to this statement [4]. A privacy policy describes the data practices of an application [5], especially what data is collected and how it is used. However, the tedious and complicated writings of the privacy policy hinder users from reading and understanding these policies [6, 7]. As a legal agreement, the natural language privacy policy lacks a machine-readable form to handle automated compliance verification, that is, whether the privacy policy provides all the information for legal requirements and the implementation complies with the privacy statements [8, 9].

There are various analysis tools to extract key information from the privacy policies to help users quickly access the policies of interest. A common practice for these tools is to classify and label the statements in the privacy policy into categories such as first-party collection and third-party sharing [10–12]. The result of rough classification is that users still need to read the statements to obtain the details,

such as the intention of the data collection. To overcome this problem, some tools introduce manual labor with specific domain knowledge to annotate the fine-grained information in the privacy policy [13], one of which requires a mean of 72 minutes per policy [5]. This solution suffers from inefficiency and is time-consuming. In terms of compliance, recent works have begun to extract structured data collections or usage statements from privacy policies to analyze data violations [14, 15]. These studies are focused on the use of undeclared data and rarely consider whether the data are used for the eligible purpose stated in policy privacy. The purpose is the key concept in data protection regulations [16]. The GDPR clearly spells out the purpose limitation; that is, "personal data shall be collected for specified, explicit, and legitimate purposes and not further processed in a manner that is incompatible with those purposes." Consequently, the expression and extraction of the purpose statement in the privacy policy are essential for data compliance [17]. However, the purpose expression is various and can be a noun or a verb phrase in natural language documents. It is not like the action element, which is naturally a verb, nor organization entities, such as "Google," "Facebook," or data objects, such as "name" and "e-mail addresses," which are types of named entities. Therefore, the extraction of purpose not only can rely on just lexical and syntactic analysis like extracting actions, entities, and data objects but also requires semantic analysis.

To overcome the above issues, we propose an approach for expressing and automatically extracting purpose-aware rules from the privacy policy, which is called PurExt. First, data collection and use statement in the privacy policy is formalized as a purpose-aware rule, which is a tuple of four elements as actor, action, data object, and purpose. The rule indicates that some *data objects* are collected or used (*action*) by a certain *actor* for a certain *purpose*. Then, through the investigation of the privacy policies, this study defines two types of purpose expressions, explicit purpose and implicit purpose, and implements the automated extraction of rule elements by combining syntactic and semantic analyses. Finally, the effectiveness of PurExt is verified on privacy policy datasets that were collected from IoT apps and manually annotated. This method can be applied in the following ways. First, this approach can be utilized to mine access control policies (ACPs) from security specifications because purpose-aware rules are nearly a draft of the ACPs. Second, PurExt can help privacy policy analysis tools to enhance the ability of fine-grained information extraction. Third, the extraction results of PurExt can be the basis and starting point for conducting data purpose compliance research.

The major contributions of this work are as follows:

(i) A purpose-aware rule is proposed to formalize and express the data collection and use statement in privacy policies. The elements of the rule are driven by the data collection and use purpose, which is the core concept of privacy protection.

(ii) The automated rule extraction of natural language privacy policies is proposed. Explicit and implicit purposes are defined to specify the purpose expression and are identified by combining SRL with syntactic analysis. To the best of our knowledge, this is the first study to extract purpose-centric rules from privacy policies. Experiments on datasets from the real IoT-related apps have proved the effectiveness of the algorithm.

(iii) Domain adaption is utilized to improve the efficiency of SRL in the field of privacy policies. The experiments show that the F1-score of the implicit purpose extraction by the domain-adapted SRL model increases 11%.

The rest of this paper is organized as follows: in Section 2, the related work in the field of privacy policy analysis in IoT, policy extraction with NLP, and SRL tools is briefly introduced. And we elaborate on the core concept of PurExt in Section 3. After that, a detailed description of PurExt is presented in Section 4. Section 5 presents the results of our extensive experimental evaluation of the proposed approach. The last section concludes the paper and points out the future research direction.

2. Related Work

2.1. Privacy Policy Analysis in IoT. By being aware of the user-unfriendly problem with respect to the privacy policy and its importance for a compliance evaluation, several studies on privacy policy analysis have emerged in recent years. Parvaneh et al. [18] utilized classification and graph-based methods to make privacy policies that are structured and categorized to help users understand them better. Onu et al. [19] defined a taxonomy framework, which uses a tree-like hierarchical form to model privacy policies within IoT environments. A framework called Polisis [20] was proposed to divide a privacy policy into fragments and assign the fragments with a set of labels that describes its data practices. Subahi and Theodorakopoulos [21] proposed eight criteria for the IoT privacy policy and implemented a test bed for ensuring the compliance of the IoT data disclosure to the corresponding privacy policy. To verify whether an application behaves according to its privacy policy, Zimmeck et al. [22] proposed using an automated analysis system that is based on machine learning and static analysis to identify potential privacy requirement inconsistencies. By considering the potential contradictions in a single privacy policy, Benjamin et al. presented an automated analysis tool called PolicyLint [14], to extract the structured data collection and sharing statements and identify contradictions among them. Furthermore, he proposed POLICHECK [23], which is an entity-sensitive flow-to-policy consistency model that is based on the results extracted from PolicyLint. Several works [24, 25] had modeled the identification of the data practice statements in privacy policies as a classification problem. The classification results that are obtained in this manner are some coarse-grained labels, which are not conducive to the compliance analysis of specific terms, which includes whether the processing of specific data adheres to the declared intention. Bhatia and Breaux [11, 26] presented a

semantic frame-based representation for data practice that can be used to identify incompleteness in a data action context. This work introduces semantic roles into the analysis of privacy policy. However, they use an inefficient way of manual annotation instead of the automatic semantic role labeling tool.

2.2. Policy Extraction with NLP. The earliest attempt to extract the data access information from the natural language security specifications is to automatically extract the ACPs with NLP. Xiao et al. [27] proposed the first work to extract ACPs from natural language software documents and produce the formal specifications, called Text2Policy. The sentences describing the ACPs were first separated from other unrelated texts by performing matching with four predefined patterns. Then, by using the annotated portions of the matched pattern, they identified the subject, action, and resource elements from the sentence. The major drawback of this approach is that ACP sentences other than the predefined patterns cannot be discerned. It has been confirmed that only 34.4% of the ACP sentences were found by matching the four patterns [28]. Nevertheless, this work is still enlightening, resulting in many follow-up studies. Slankas and Williams [29] proposed the concept of access control relation extraction (ACRE), which is a method of incorporating machine learning and NLP to extract ACP elements. They used classification algorithms to determine whether these sentences are related to the access control. A bootstrapping process was adopted to extract the ACP instance from a small set of seeded dependency graph patterns. Subsequently, an extended ACRE was proposed [28]. Unlike the previous approach, the votes for the K-nearest neighbor (KNN), naive Bayes, and simple vector machine were replaced by the KNN classification algorithm in the sentence identification phase. Narouei et al. [30] introduced four different types of features to improve the effect of distinguishing the ACP sentences from other sentences. Subsequently, they tried to use semantic role labeling to identify ACP elements [31], but the identification effects of this method on different data sets are quite different. All the above methods can be used for policy extraction. However, their extracted objects are the elements of ACPs, such as roles, resources, and attributes and barely involve data collection and use purpose, which is an important concept in the privacy policy.

2.3. SRL Tools. SRL is a shallow semantic parsing task, in which the goal is to identify the arguments of the verb predicate in a sentence and assign semantic labels to those arguments [32]. SRL starts with the action predicate of a sentence to determine other sentence constituents that correspond to who did what to whom, when, where, and why. This information is useful for identifying the key concepts of the data collection and usage statements from a sentence. To evaluate the effects of different SRL tools in extracting the structured data collection and usage statements, the PurExt algorithm is implemented based on four SRL tools. EasySRL [33], which is written in Java, provides a semantic role labeler and combinatory categorial grammar parser. Mate-tools [34] provide a pipeline of modules that perform the lemmatization, part-of-speech (POS) tagging, dependency parsing, and SRL of a sentence. The tools are language-independent and have high accuracy. Semantic/syntactic extraction using a neural network architecture (SENNA) [35] is a SRL program that is trained on a PropBank corpus, which also offers other common NLP tasks such as POS tagging, chunking, and named entity recognition (NER). Unlike other SRL systems, SENNA assigns semantic roles to sentence constituents without the help of a syntax tree, thus resulting in better efficiency. SwiRL [36] is a SRL system for English that is constructed on top of the full syntactic analysis of the text. SwiRL has a user-friendly feature; that is, the model can be retrained through the application programming interface (API) that is provided by the system, which enables the user to add domain-specific knowledge.

3. Purpose-Aware Rule

In this chapter, we define the purpose-aware rule based on the meaning of purpose in the privacy policy, which specifies the objects to be extracted. And the purpose expression in natural language is analyzed to clarify the extraction method.

3.1. Purpose in Privacy Policy. Purpose is the key concept in privacy policy, which explains the reason for data collection and use. It determines whether the user will agree to the data collection and use behavior of an application. Although the purpose is shown as a few words or phrases in the privacy policy, the connotation of purpose is a constraint on what data is collected and how it is used.

Example 1. We illustrate the meaning of purpose using the privacy policy from Xiaomi Wear App. A statement of "to facilitate the registration of your smart wearables in the app, we may collect the information related to your Mi Account, identifier of smart wearables, identifier of your phone (IMEI number encrypted via Hash algorithm), phone model, OS version, and Bluetooth information of smart wearables" in the privacy policy indicates that the data such as account, identifier of device, and OS version is only allowed to be collected for the purpose of the registration of smart wearables, as shown in Figure 1.

From the example, we can see that a purpose is specific to the related data objects, the action performed on the data, and the actor that performs the action. Driven by this insight, a data collection and use statement is formalized as a purpose-aware rule, which can be regarded as a collection of actors, actions, and data objects driven by a purpose.

Definition 1 (purpose-aware rule). *A purpose-aware rule is defined as a tuple:*

$$\text{rule} = \{\text{actor}, \text{action}, \text{data object}, \text{purpose}\}, \quad (1)$$

where *actor* is the entity that performs the action on a data object, *action* is the operation performed on the data objects,

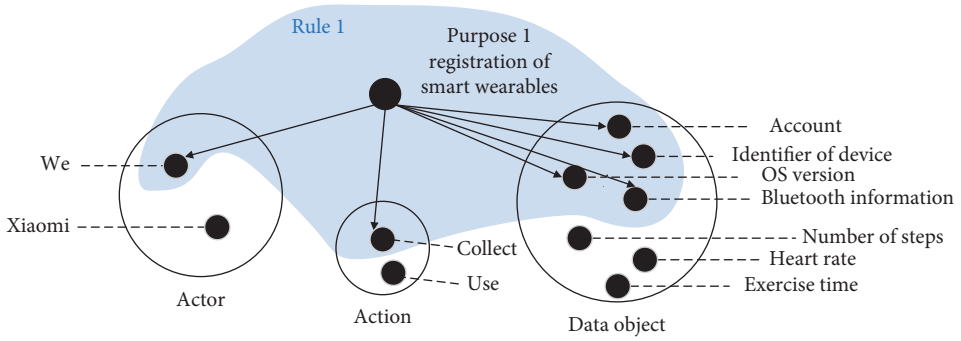

Figure 1: An example of the purpose constraint.

data object is the data related to the purpose, and *purpose* is the intention of data collection and usage.

Example 2. The statement in Example 1 can be formalized as a purpose-aware rule like r_1({we}, {collect}, {Mi Account, identifier of smart wearables, identifier of your phone (IMEI number encrypted via Hash algorithm), phone model, OS version, Bluetooth information of smart wearables}, the registration of your smart wearables in the app).

A benefit of the rule is the direct or internal connection between its elements and the concepts of data access control at the implementation layer. Although the rule elements extracted from the privacy policies are mostly abstract, the high-level user-oriented rule can be refined into low-level operational policies via hierarchical and deductive methods [37, 38]. For example, the data object "personal information" that most frequently appears in the privacy policy can be refined and mapped into fields such as "name" and "age" in the database. On the contrary, it is also possible to judge whether the underlying data practice complies with the data use statement in the privacy policy through induction of elements.

3.2. Purpose Expression in Natural Language. Concerning the expression of the rule elements in natural language, actor and data object are usually nouns and regarded as a type of named entity [14], and action is present as a verb. Consequently, they can be identified and extracted from the natural language sentences based on these lexical features. However, the appearance of purpose in sentences is ambiguous.

Because the purpose is to explain the reason for the data collection and use behaviors, it is usually acted as the semantic role of the purpose attach with some keywords, such as "in order to" and "for," to modify the verbs of collection and use action in the sentence; for example, "we use purchase records in order to analyze user preferences." A statistic shows that 78% of the usage actions in the privacy policy are modified by the purpose semantic role [26]. According to our observations on a large number of privacy policies, this is also the most common way in which the purpose appears in the privacy policy document. Meanwhile, we found another appearance of purpose like "the purpose of collecting purchase records is preference analysis." In this case, the predicate of the sentence is no longer a collection or use (CoU) verb, so the purpose cannot be attached to the predicate as a semantic role but can only appear independently. Through analyzing the structure and key constituents of the two kinds of sentences, two patterns of purpose expression in natural language privacy policies are concluded as follows:

3.2.1. Explicit sentence. The backbone of this kind of sentence is to state the purpose straightforwardly. The pattern is $P_{\text{noun}} + V_{\text{link}}/V_{\text{contain}} + \text{Purpose}$, where P_{noun} is the noun representing purpose, V_{link} is the linking verb such as "is" and "are," and V_{contain} is the verb representing the meaning of containing.

Example 3. "The purpose of collecting your location data and speed is to analyze your train statistics" and "the reason of using your name and e-mail address includes website account registration" are explicit sentences.

Implicit sentence: the backbone of this kind of sentence is to state the data collection or use behavior, and the collection or use action is modified by the purpose semantic role. The pattern is

$$sbj + V_{CoU} + \text{Data} + \text{Purpose}, \tag{2}$$

where V_{CoU} is the verb representing collection and use actions, *sbj* is the subject of the V_{CoU}, and *Data* represents the collected and used data objects.

Example 4. "The app will collect your heart rate and pulse to make suggestions for future workout" is an implicit sentence.

The two patterns only represent different manifestations of the purpose in the natural language sentences, but the semantics of the purpose will not differ depending on its manifestation. The definition and feature of the two patterns also determine the extraction methods for the two purposes of them. The purpose in the implicit sentence can be readily identified using semantic role labeling, while the purpose in the explicit sentence can be extracted based on the syntactic features of the pattern. To distinguish the source of the extracted purpose, the purpose extracted from an explicit sentence is called the *explicit purpose*, and the purpose from an implicit sentence is called the *implicit purpose*.

4. PurExt

In order to identify the purpose-aware rules from the privacy policies, we propose an automated rule extraction framework, PurExt. An overall view of the proposed approach is shown in Figure 2. Details of each step are described in the following sections.

4.1. Preprocessing. This study used the open-source library, spaCy [39], to perform a series of essential natural language preprocessing on sentences. SpaCy is a fast NLP toolkit that is implemented in Python. The tokenization, POS tagging, and NER of a sentence can be easily and serially completed by constructing a pipeline task. The tokenization splits the text into individual words, which allows us to obtain the smallest unit of processing. The POS tag that is assigned to the token provides the lexical information that is needed for NER and dependency parsing. As an important preliminary task of relationship extraction, the goal of NER is to identify phrases that represent real-world objects, such as the name, geographic location, organization, and date, and label them with corresponding named entity tags. PurExt applies a domain-adapted NER model to label the entities concerning the field of the privacy policy. For example, "name" and "age" are annotated as the named entity type of *Data Object*, and "we" and "advertisers" are annotated as the named entity type of *Entity*. The NER model has the ability to identify *Data Object* and *Entity* that are specific to privacy policies by retraining spaCy's NER engine with 500 annotated sentences from the privacy policies.

4.2. Sentence Classification. The goal of sentence classification is to divide the sentences into explicit sentences, implicit sentences, and other sentences. Both explicit sentences and implicit sentences are related to the data collection or use statement, as well as the target of the rule extraction, so we call them CoU sentences for brevity. Other sentences describe the information such as terms of service, data retention, policy updates, or contact information. Because they have nothing to do with the data collection or use statement, they are not processed.

As described in Algorithm 1 SentenceClassification (), the first step of sentence classification is to build a dependency tree for each sentence. The parsed dependency tree sketches the lexical structure of a sentence by constructing a dependency relationship between words.

Sentence classification starts from the root node of the dependency tree. Explicit sentences have two methods of stating the purpose. (1) The predicate of the sentence indicates the relationship of existence, which is usually a linking verb (denoted as V_{link}), meaning "what is the purpose." (2) The predicate indicates the containment relationship, which is usually a verb that represents the meaning of containing (denoted as $V_{contain}$), such as "include" and "contain," which conveys "what does the purpose include." Therefore, if the predicate of the sentence satisfies one of the conditions, the sentence has the potential to be an explicit sentence. However, if the predicate describes the data collection or use behaviors (for brevity, these verbs are called CoU verbs, which are listed in Table 1), the sentence is a potential implicit sentence. Those that do not meet the above two conditions are classified as other sentences and will not be analyzed. Candidate sentences that are filtered by the predicate are identified as explicit sentences in two ways. First, the subject of its predicate is the purpose noun (called Pnoun for brevity, which is listed in Table 1), that is, a noun that refers to the purpose. Second, the Pnoun is modified by a complement that describes the data collection or use statement. This constraint aims to avoid the wrong purpose affiliation. The additional condition of the implicit sentence is to contain at least one *Data Object*. Through sentence classification, the processing range is narrowed, and the location of the elements that is to be extracted is roughly located.

4.3. Rule Extraction. The goal of this step is to extract the actor, action, data object, and purpose from the sentences to form the data security rules. Because of the different expressions of the purpose, the element extraction of explicit sentences is based on the syntactic analysis, whereas the extraction of explicit sentences is mainly based on the semantic analysis.

4.3.1. Element Extraction of Explicit Sentences. According to the definition of the explicit sentence, the subject of the predicate, that is, Pnoun, is modified by the data collection or use statement, which indicates the affiliation of the purpose. Therefore, the extraction of the action and data object begins by parsing the prepositional complement of the Pnoun. PurExt traverses down the branch to identify the node matching CoU verbs as an action element and extracts all the nodes that are annotated as data objects. The purpose expression in explicit sentences is derived into verbs and nouns. The first one is usually used as a complement to modify the predicate, while the other appears in the form of an object. The verbal purpose expression is associated with the predicate by the "xcomp" dependency label, which means that it is the open clause complement of the predicate. PurExt parses this branch and forms a purpose phrase. As for the purpose expression in the noun form, PurExt will parse the branch, linking to the predicate with a "dobj" (direct object) dependency tag or a "pobj" (object of a preposition) tag, beginning with a preposition. This is because there is a case where the purpose phrase and predicate are connected by a preposition; for example, "the purpose of collecting your personal information is for legal obligations." Actor elements are usually not involved in explicit sentences.

Example 5. The two explicit sentences in Example 3 have the purpose of verb form and noun form, respectively. The dependency trees of the sentences are shown in Figure 3. And the two sentences can be extracted as r_1({}, {collect}, {your location data, speed}, analyze your train statistics), and r_2({}, {use}, {your name, e-mail address}, website account registration).

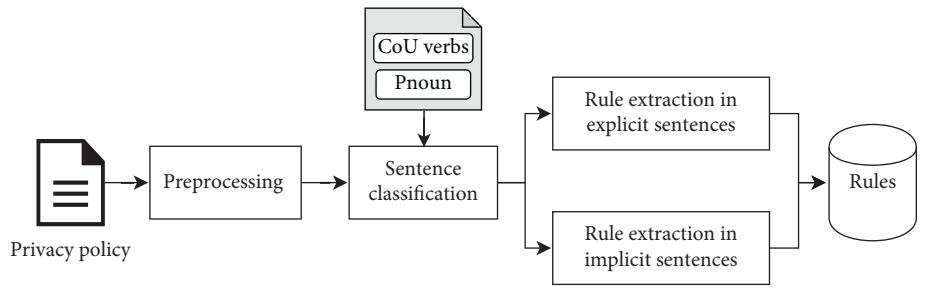

Figure 2: Overview of PurExt.

Input: a sentence S to be classified.
Output: a sentence category label in (E, I, O), where E for explicit sentence, I for implicit sentence, and O for other sentences.

(1) Construct a dependency tree structure T of S
(2) Let p be the root of T
(3) **if** p is a V_{link} or $V_{contain}$ **then**
(4) **if** the subject s of p is a Pnoun **then**
(5) **if** s is modified by a complement containing at least one CoU verb **then**
(6) **return** E.
(7) **else**
(8) **return** O
(9) **end if**
(10) **else**
(11) **return** O
(12) **end if**
(13) **else if** p is CoU verb **then**
(14) **if** the object of p contains at least one *Data Object* **then**
(15) **return** I
(16) **else**
(17) **return** O
(18) **end if**
(19) **else**
(20) **return** O
(21) **end if**

Algorithm 1: SentenceClassification ().

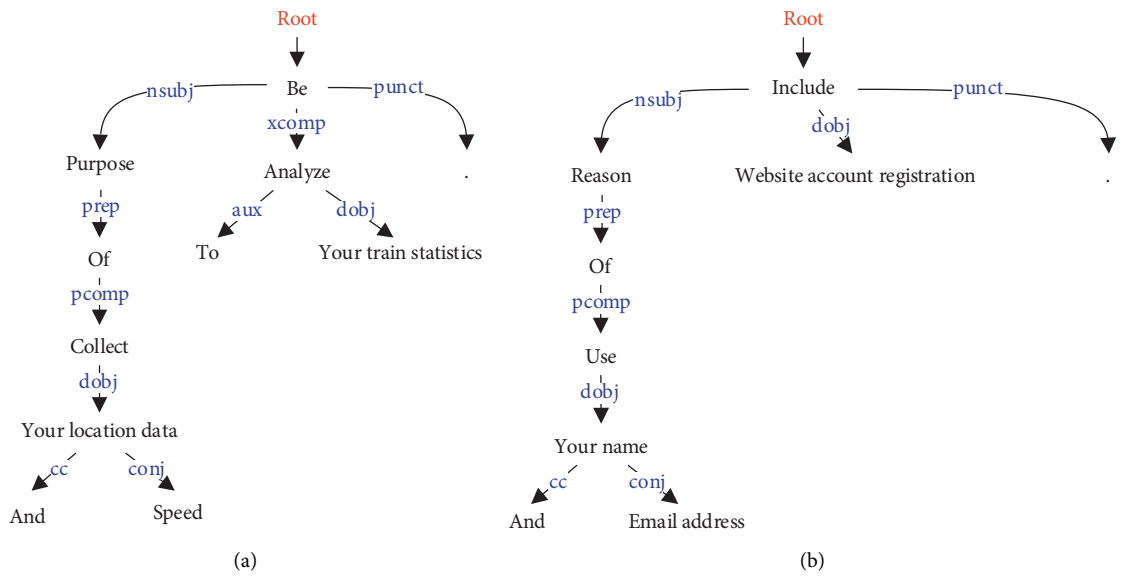

Figure 3: Dependency tree of the explicit sentence whose purpose representation is in (a) verb form and (b) noun form.

4.3.2. Element Extraction of Implicit Sentences. The purpose phrase is not the main component in implicit sentences. As the role of modifying other ingredients, it has various forms. Semantic analysis can cross the different syntactic structures of a sentence to present its semantic information. For example, although the grammatical construction of "A car hit Bob" and "Bob was hit by a car" are different, the semantic information for them is the same and can be unified as hit (car, Bob). Therefore, PurExt adopts SRL [40], a shallow semantic parsing, to infer the purpose of the data collection or use statements in implicit sentences. The SRL detects the semantic arguments that are related to the verbs in a sentence, to assign semantic roles that are defined by the PropBank project to the constituents of the sentence. The meaning of common semantic roles is shown as follows:

V: verb.

A0: agent of the verb.

A1: patient.

AM-MOD: modal.

AM-PNC: purpose.

The connotation of a data security rule is that an actor performs collection or use actions on data objects for a certain purpose. Reasonably, the actor corresponds to the "A0″ role of the CoU verbs, the data objects match the "A1″ role, and the purpose is the "AM-PNC" role. Hence, through this mapping, four elements of a rule can be extracted with the help of SRL. To improve the accuracy of the extraction, the constraint of the named entity type is also attached. In addition to being marked as "A0," the actor element needs to be marked as *Entity* by NER. Similarly, data objects should be annotated as the "A1″ role, as well as the entity type of *Data Object*.

Example 6. The sentence in Example 4 is annotated as "[The app **A0**] [will **AM-MOD**] [collect **V**] [your heart rate and pulse **A1**] [to make suggestions for future workout **AM-PNC**]" by the SRL tool. And the rule extracted from the sentence is r({the app}, {collect}, {your heart rate, pulse}, make suggestions for future workout).

4.4. Domain Adaption. The success of the extraction primarily depends on the accurate annotation of the SRL. By considering the labeling effect and user-friendliness, we choose SwiRL to implement PurExt. SwiRL is a SRL tool that is trained on the PropBank corpus, which is a collection of annotated sentences from the Wall Street Journal. The terms and expressions in financial magazines are different from those in privacy policies. Therefore, the model that migrates into this problem domain is prone to mislabeling. For example, the purpose phrase that should be labeled as "AM-PNC" is incorrectly assigned to the "A2" role, which leads to false negative results. This is also confirmed by the experiment in Section 5.3.

To mitigate this problem, this study tried to adapt the SRL tool SwiRL to the privacy policy domain; that is, we use a small number of domain-specific datasets to retrain SwiRl, such that it can obtain more accurate annotation results in the target domain. For the preparation, 400 purpose-related sentences of privacy policies are collected from real apps and manually marked with the semantic roles. The experiment in Section 5.4 proved that the effect of implicit purpose extraction, which uses SRL annotation, improves significantly after the domain adaption.

5. Experiments

In this section, we present three evaluations conducted to assess the effectiveness of our proposed approach. In our evaluations, we specifically focus on the following questions:

RQ1: how effectively does PurExt extract the explicit purpose and implicit purpose?

RQ2: is there any improvement in PurExt with domain adaption applied?

RQ3: how *effectively* does PurExt extract the other three elements, that is, actor, action, and data object?

5.1. Dataset. Because of the lack of a public dataset annotated with the purpose statement, we collected and annotated a dataset from the privacy policies of real APPs, which mainly belong to the Wear OS, healthcare, and other IoT-related APP categories in Google Play Store. First, the privacy policy HTML files of each app were downloaded with a crawler program developed by authors. Then, a tool named HtmlToPlaintext [41] was used to convert these HTML files into plaintext policy documents. Finally, a doctoral student and a master student with background knowledge selected 1,000 sentences from these documents and annotated the phrase in each sentence with the element labels that they related to.

There are 750 CoU sentences, which consist of 584 implicit sentences with purpose statement, 46 without purpose, and 120 explicit sentences. In order to align the distribution of the dataset with the real privacy policy, there are 250 sentences describing another nine types of data practice [5], such as third-party sharing/collection, data retention, data security, and policy change. To perform and verify the domain adaptation experiment, about 68 percent of the 584 implicit sentences were used for training and 32 percent for testing. SwiRL model was retrained on the dataset composed of the 400 implicit sentences and its original training corpus. Apart from the 400 sentences, the remaining 600 sentences constitute the test dataset. Table 2 summarizes the number of different sentences along with their annotated elements in the test dataset.

5.2. Evaluation Criteria. To assess the effectiveness of the element extraction, the experimental results were measured with respect to the precision (P), recall (R), and F1-score (F1) [42], which are defined as follows: To compute these values, the experimental results are divided into four

Table 1: Word lists.

Type	Word
CoU verbs	Access, check, collect, disclose, gather, keep, know, obtain, process, provide, receive, request, retain, save, share, store, transfer, update, use, utilize
Pnoun	Purpose, reason, intention, goal, motivation, way

categories. True positive (TP) means extracting the element correctly. False positive (FP) represents a case in which an unrelated constituent is identified as an element. False negative (FN) is the prediction that a true element in a sentence is not recognized. Finally, true negative (TN) is a situation in which the approach correctly identifies a constituent to be unrelated:

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}},$$
$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \quad (3)$$
$$F1 = 2 * \frac{P * R}{P + R}.$$

5.3. Effect of Purpose Extraction

5.3.1. Experiment Setup. We separately counted the TP, FP, and FN values for the explicit and implicit purposes, to show the extraction effect of PurExt for the two types of purposes. To determine how different SRL tools affect implicit purpose extraction, PurExt is implemented based on the semantic annotations from four different SRL systems. The four SRL systems are EasySRL, Mate-tools, SENNA, and SwiRL, which were introduced in Section 2.3. Because explicit purpose extraction does not use SRL, the explicit purpose extraction results of the four SRL tools are the same.

5.3.2. Experiment Results. The result of the explicit purpose extraction is shown in Table 3. The position of the explicit purpose in the sentence is relatively fixed; thus, the precision of the extraction based on the syntactic structure is high. However, the recall rate for explicit purposes is comparatively lower. We analyzed the FNs in the results and found that the sentences, which were not detected by PurExt, do not have a clause modification of the subject that describes the data usage behavior, which is one of the criteria for an explicit sentence. For example, consider "the purpose is to process your payment." The privacy policy document is rich in context; hence, the subordination of the purpose may appear in the title or elsewhere. This method represents sentence-level NLP, which will misjudge the sentences that are separated from their dependency.

Table 4 shows that the purpose identified from the sentences with four SRL tools did not reach half of the total. By digging into the annotations of the four SRL tools, we observed a phenomenon that several purpose phrases, which should be marked as the *AM-PNC* role, are assigned as the *A*2 role. As mentioned in Section 4.4, these tools are trained on another domain corpus, such as PropBank and FrameNet. Because of the difference in word usage and expression habits, they do not perform well in assigning semantic labels for the privacy policy documents.

5.4. Effect of Domain Adaption

5.4.1. Experiment Setup. Domain adaption is introduced to improve the recall rate of implicit purpose extraction via SRL. Considering the trainability of the tools and the effect that is shown in the first experiment, SwiRL, which is user-friendly and provides a retraining API, was chosen as the target. The training dataset is described in Section 4.4. This section demonstrates the overall effect of PurExt that is achieved with the domain-adapted SwiRL and a separate result of the implicit purpose extraction, which can directly reflect the effect of the domain adaption.

5.4.2. Experiment Results. Table 5 shows that retraining with a small amount of domain-specific data significantly improves the effect of SRL on implicit purpose extraction. In detail, the recall rate and F1-score increase by almost 13% and 11%, respectively. In the future, we will continue to explore the effect of the training dataset size on the retrained model. The overall extraction results of PurExt that are realized by the domain-adapted SwiRL are shown in Table 6. The precision of each element extraction is over 90%, and the precision of the rule reaches 97%. In terms of the recall rate, except for the purpose, which is 69%, the other elements are all over 85%. The F1-score for the rule reaches 91%.

5.5. Effect of Other Elements Extraction

5.5.1. Experiment Setup. To determine how effectively PurExt extracts the elements of the actors, actions, and data objects, PurExt is compared with a recent work PolicyLint [14] that is consistent with the targets that were extracted from the privacy policies in this investigation. PolicyLint is a privacy policy analysis tool that identifies potential contradictions that may arise inside the same privacy policy. It provides a sentence-level NLP method to capture sharing and collection statements in privacy policies as a four-tuple (actor, action, data object, entity), where entity corresponds to the object of data sharing. We run PolicyLint on the privacy policy dataset and compare its extraction results, just the three-tuple (actor, action, data object), with the results of our approach to evaluate the efficiency of PurExt on the extraction of actor, action, and data object.

Table 2: Test dataset statistics.

		#sentence	#actor	#action	#data object	#purpose
CoU sentences	Explicit	120	0	112	107	120
	Implicit	184	125	184	362	184
	Other	46	45	46	205	0
Unrelated sentences	250	0	0	0	0	

Table 3: Effect of explicit purpose extraction.

	P (%)	*R* (%)	F1 (%)
Explicit purpose	100.00	82.50%	90.41%

Table 4: Effect of implicit purpose extraction.

	Implicit purpose		
	P (%)	*R* (%)	F1 (%)
EasySRL	100.00	29.89	46.03
Mate-tools	97.75	47.28	63.74
SENNA	98.91	49.46	65.94
SwiRL	97.75	47.28	63.74

Table 5: Effect of domain adaption on implicit purpose extraction.

	Implicit purpose		
	P (%)	*R* (%)	F1 (%)
SwiRL	97.78	47.57	64.00
SwiRL_DA	99.12	60.54	75.17

Table 6: The overall effect of PurExt.

	P (%)	*R* (%)	F1 (%)
Actor	92.45	85.47	88.82
Action	96.18	95.06	95.61
Data	97.90	89.91	93.74
Purpose	99.53	69.08	81.55
Rule	97.07	86.35	91.39

Table 7: Effect of other elements extraction.

	Actor			Action			Data		
	P (%)	*R* (%)	F1 (%)	*P* (%)	*R* (%)	F1 (%)	*P* (%)	*R* (%)	F1 (%)
PolicyLint	74.86	77.91	76.35	90.50	69.53	78.64	95.04	57.50	71.65
PurExt	92.45	85.47	88.82	94.32	92.70	93.51	97.48	88.71	92.89

5.5.2. Experiment Results. Table 7 demonstrates the performance of PolicyLint and PurExt to extract three elements. As demonstrated, PurExt performs better than PolicyLint in all aspects. On the one hand, the advantage of PurExt is that it benefits from the combination of syntactic and semantic analyses compared to PolicyLint, which only uses the former. On the other hand, PolicyLint uses 82 templates that were learned from 560 example sentences to match the sentences to be extracted; thus, it passes over the sentences that do not follow the patterns.

6. Conclusions

Because IoT devices collect a considerable amount of personal and sensitive information, the privacy issues for IoT are a major concern for the users and laws. Privacy policy is an important way for IoT vendors to obtain users' trust and to adhere to legal requirements. Therefore, how to make the privacy policy better serve users and regulatory compliance has aroused our interest. This study explores the expression and automated extraction of the purpose-centric data usage

purposes in privacy policies. More precisely, we propose a purpose-aware rule to formalize the data access statements and combine syntactic and semantic analyses to realize the automated extraction of rules from the natural language privacy policies. To the best of our knowledge, this is the first attempt to extract the structural purpose-centric statement from privacy policies. The experimental results on real datasets show that this approach can achieve a 91% recall rate and 97% precision.

Because PurExt performs sentence-level extraction, a separate description of the data objects and purposes in different sentences will lead to incomplete rules. When considering the future directions of research, the entire document should be analyzed to obtain more context. In terms of the promotion, we will proceed to apply the extracted purpose-aware rules for the research of compliance verification.

Data Availability

The data used to support the findings of this study are available from the first author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (nos. U19A2081 and 61802270).

References

[1] C. B. Foltz and L. Foltz, "Mobile users' information privacy concerns instrument and IoT," *Information & Computer Security*, vol. 28, no. 3, pp. 359–371, 2020.

[2] E. Zeng and F. Roesner, "Understanding and improving security and privacy in multi-user smart homes: a design exploration and in-home user study," in *Proceedings of the 28th USENIX Security Symposium (USENIX Security 19)*, pp. 159–176, Berkeley, CA, USA, May 2019.

[3] S. Zheng, N. Apthorpe, M. Chetty, and N. Feamster, "User perceptions of smart home IoT privacy," *Proceedings of the ACM on Human-Computer Interaction*, vol. 2, pp. 1–20, 2018.

[4] T. Pasquier, J. Singh, J. Powles, D. Eyers, M. Seltzer, and J. Bacon, "Data provenance to audit compliance with privacy policy in the Internet of Things," *Personal and Ubiquitous Computing*, vol. 22, no. 2, pp. 333–344, 2018.

[5] S. Wilson, F. Schaub, A. A. Dara et al., "The creation and analysis of a website privacy policy corpus," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 1330–1340, Berlin, Germany, August 2016.

[6] J. Mohan, M. Wasserman, and V. Chidambaram, "Analyzing gdpr compliance through the lens of privacy policy," in *Heterogeneous Data Management, Polystores, and Analytics for Healthc*, pp. 82–95, Springer, Berlin, Germany, 2019.

[7] W. B. Tesfay, P. Hofmann, T. Nakamura, S. Kiyomoto, and J. Serna, "I read but don't agree: privacy policy benchmarking using machine learning and the EU GDPR," in *Proceedings of the Web Conference 2018 - Companion of the World Wide Web Conference, WWW 2018, Association for Computing Machinery, Inc*, pp. 163–166, Lyon, France, April 2018.

[8] G. Kapitsaki, J. Ioannou, J. Cardoso, and C. Pedrinaci, "Linked USDL privacy: describing privacy policies for services," in *Proceedings of the 2018 IEEE International Conference on Web Services (ICWS)*, pp. 50–57, IEEE, Seattle, WA, USA, June 2018.

[9] R. N. Zaeem, R. L. German, and K. S. Barber, "Privacycheck: automatic summarization of privacy policies using data mining," *ACM Transactions on Internet Technology*, vol. 18, no. 4, pp. 1–18, 2018.

[10] D. A. Audich, R. Dara, and B. Nonnecke, "Privacy policy annotation for semi-automated analysis: a cost-effective approach," in *IFIP Advances in Information and Communication Technology*, pp. 29–44, Springer, Berlin, Germany, 2018.

[11] J. Bhatia and T. D. Breaux, "Semantic incompleteness in privacy policy goals," in *Proceddings of the 2018 IEEE 26th International Requirements Engineering Conference (RE)*, pp. 159–169, IEEE, Banff, AB, Canada, August 2018.

[12] N. M. Nejad, P. Jabat, R. Nedelchev, S. Scerri, and D. Graux, "Establishing a strong baseline for privacy policy classification," in *IFIP International Conference on ICT Systems Security and Privacy Protection*, pp. 370–383, Springer, Berlin, Germany, 2020.

[13] S. Wilson, F. Schaub, R. Ramanath et al., "Crowdsourcing annotations for websites' privacy policies: can it really work?" in *Proceedings of the 25th International Conference on World Wide Web*, pp. 133–143, Montreal, Canada, April 2016.

[14] B. Andow, S. Y. Mahmud, W. Wang, J. Whitaker, W. Enck, B. Reaves et al., "Policylint: investigating internal privacy policy contradictions on google play," in *Proceedings of the 28th USENIX Security Symposium (USENIXSecurity 19)*, pp. 585–602, Berkeley, CA, USA, August 2019.

[15] L. Yu, X. Luo, C. Qian, S. Wang, and H. K. Leung, "Enhancing the description-to-behavior fidelity in android apps with privacy policy," *IEEE Transactions on Software Engineering*, vol. 44, no. 9, pp. 834–854, 2017.

[16] M. C. Tschantz, A. Datta, and J. M. Wing, "Purpose restrictions on information use," in *Lecture Notes in Computer Science*, pp. 610–627, Springer, Berlin, Germany, 2013.

[17] D. Basin, S. Debois, and T. Hildebrandt, "On purpose and by necessity: compliance under the GDPR," in *Financial Cryptography and Data Security*, pp. 20–37, Springer, Berlin, Germany, 2018.

[18] P. Shayegh, V. Jain, A. Rabinia, and S. Ghanavati, "Automated approach to improve iot privacy policies," 2019, http://arxiv.org/abs/1910.04133.

[19] E. Onu, M. M. Kwakye, and K. Barker, "Contextual privacy policy modeling in iot," in *Proceedings of the 2020 IEEE International Conference on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing*, pp. 94–102, Calgary, AB, Canada, August 2020.

[20] H. Harkous, K. Fawaz, R. Lebret, F. Schaub, K. G. Shin, and K. Aberer, "Polisis: automated analysis and presentation of privacy policies using deep learning," in *Proceedings of the 27th USENIX Security Symposium (USENIX Security 18)*, pp. 531–548, Berkeley, CA, USA, August 2018.

[21] A. Subahi and G. Theodorakopoulos, "Ensuring compliance of IoT devices with their privacy policy agreement," in *Proceedings of the 2018 IEEE 6th International Conference on Future Internet of Things and Cloud (FiCloud)*, pp. 100–107, Barcelona, Spain, August 2018.

[22] S. Zimmeck, Z. Wang, L. Zou et al., "Automated analysis of privacy requirements for mobile apps," in *Proceedings of the NDSS*, San Diego, CA, USA, May 2017.

[23] B. Andow, S. Y. Mahmud, J. Whitaker, W. Enck et al., "Actions speak louder than words: entity-sensitive privacy policy and data flow analysis with policheck," in *Proceedings of the 29th USENIX Security Symposium (USENIX Security 20)*, pp. 985–1002, Anaheim, CA, USA, August 2020.

[24] P. Story, S. Zimmeck, A. Ravichander et al., "Natural language processing for mobile app privacy compliance," in *Proceedings of the AAAI Spring Symposium on Privacy-Enhancing Artificial Intelligence and Language Technologies*, PaloAlto, CA, USA, March 2019.

[25] S. Zimmeck, P. Story, D. Smullen et al., "Maps: scaling privacy compliance analysis to a million apps," *Proceedings on Privacy Enhancing Technologies*, vol. 2019, no. 3, pp. 66–86, 2019.

[26] J. Bhatia, M. C. Evans, and T. D. Breaux, "Identifying incompleteness in privacy policy goals using semantic frames," *Requirements Engineering*, vol. 24, no. 3, pp. 291–313, 2019.

[27] X. Xiao, A. Paradkar, S. Thummalapenta, and T. Xie, "Automated extraction of security policies from natural-language software documents," in *Proceedings of the ACM SIGSOFT 20th International Symposium on the Foundations of Software Engineering*, pp. 1–11, Cary, NC, USA, November 2012.

[28] J. Slankas, X. Xiao, L. Williams, and T. Xie, "Relation extraction for inferring access control rules from natural language artifacts," in *Proceedings of the 30th Annual Computer Security Applications Conference*, pp. 366–375, NewYork, NY, USA, December 2014.

[29] J. Slankas and L. Williams, "Access control policy extraction from unconstrained natural language text," in *Procedings of the 2013 International Conference on Social Computing*, pp. 435–440, IEEE, Alexandria, VA, USA, September 2013.

[30] M. Narouei, H. Khanpour, and H. Takabi, "Identification of access control policy sentences from natural language policy documents," in *Data and Applications Security and Privacy XXXI*, pp. 82–100, Springer, Berlin, Germany, 2017.

[31] M. Narouei, H. Takabi, and R. Nielsen, "Automatic extraction of access control policies from natural language documents," *IEEE Transactions on Dependable and Secure Computing*, vol. 17, no. 3, pp. 506–517, 2018.

[32] V. Punyakanok, D. Roth, and W.-t. Yih, "The importance of syntactic parsing and inference in semantic role labeling," *Computational Linguistics*, vol. 34, no. 2, pp. 257–287, 2008.

[33] M. Lewis, L. He, and L. Zettlemoyer, "Joint a* ccg parsing and semantic role labelling," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing,*, pp. 1444–1454.

[34] A. Björkelund, L. Hafdell, and P. Nugues, "Multilingual semantic role labeling," in *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pp. 43–48, Boulder, CO, USA, June 2009.

[35] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of Machine Learning Research*, vol. 12, pp. 2493–2537, 2011.

[36] M. Surdeanu and J. Turmo, "Semantic role labeling using complete syntactic analysis," in *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pp. 221–224, 2005.

[37] A. I. Antón, E. Bertino, N. Li, and T. Yu, "A roadmap for comprehensive online privacy policy management," *Communications of the ACM*, vol. 50, no. 7, pp. 109–116, 2007.

[38] X. Yang and J. Alves-Foss, "Security policy refinement: high-level specification to low-level implementation," in *Proceedings of the 2013 International Conference on Social Computing*, pp. 502–511, IEEE, Washington, DC, USA, September 2013.

[39] M. Honnibal and I. Montani, "Spacy 2: Natural language understanding with bloom embeddings," *Convolutional Neural Networks and Incremental Parsing*, vol. 7, no. 1, 2017.

[40] X. Carreras and L. Màrquez, "Introduction to the CoNLL-2005 shared task: semantic role labeling," in *Proceedings of the ninth conference on computational natural language learning (CoNLL-2005)*, pp. 152–164, Ann Arbor, MI, USA, June 2005.

[41] Github Htmltoplaintext. https://github.com/benandow/HtmlToPlaintext.

[42] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.